

United States Patent Application

Title of the Invention

A METHOD OF SEARCHING DOCUMENTS AND
A SERVICE FOR SEARCHING DOCUMENTS

Inventors

Makoto IWAYAMA,

Shingo NISHIOKA,

Yoshiki NIWA,

Akihiko TAKANO.

Title of the Invention

A method of searching documents and a service for
searching documents

5 Background of the Invention

This invention relates to a document searching
method for changing over between plural document
databases, and constructing relationships between
plural document databases.

10 As more and more document information is
converted to electronic format, a greater need is
emerging to search different types of document database
simultaneously. For instance, users often wish to look
up dictionaries relating to newspaper articles which
15 they may find of interest.

In the past, it was possible to perform a
search independently by changing over between plural
document databases, but there was no way of examining
the relevancy of sets of documents in other databases
20 to a set of documents in one particular database.

If however the search is limited to the same
document database, it is possible to search other
document sets within that database. In this case,
sufficient search speed is often obtained by
25 calculating the relevance between documents before
searching. Even with different databases, it is
possible to search plural document databases at the
same time if such a calculation is performed beforehand,

but since the need for calculation increases as the number of databases increases due to increasing numbers of combinations, this method is not realistic.

It is also possible to first analyze the set of key documents on the user side to compose a search input, and then search in other document databases by using the input, but in this case, the user side has to receive all the information about the set of key documents, and if the document databases are on a network, the amount of traffic would be huge.

Summary of the Invention

It is therefore an object of this invention to resolve the problems inherent in existing technology by allowing a user to specify an arbitrary set of documents in an arbitrary document database, and to efficiently search sets of documents relating to this set of documents from within any particular database.

When there is a large search input as in the case of a set of key documents, instead of using all the information in the search input, it is faster to perform a search using only topic words of the search input as a summary, and this also reduces the load on the network. In the context of this specification, "summary" means a "set of topic words for a set of documents".

The document databases are located on servers on a network comprising a module for building a

summary by selecting topic words for a set of documents within the document database, and a module for performing a search on any arbitrary summary.

5 A user who performs a search specifies a set of documents via a client to a server in which an source document database is stored, and receives a summary.

Next, the summary is sent to a server where a target document database to be searched is located, and a search result is received.

10 As the search interface of the client, a display area for a set of documents is first provided wherein the required set of key documents can be specified, and the database to be searched can also be selected. In the client, the user then selects an
15 interesting set of documents from among a set of documents displayed in this display area, and if necessary, changes over the document database which is to be searched.

20 These and other objects, features and advantages of the present invention will become more apparent in view of the following detailed description of the preferred embodiments in conjunction with the accompanying drawings.

25 Brief Description of the Drawings

Fig. 1 is a diagram showing an example of the overall construction of a system implementing the plural document database search method.

Fig. 2 is a diagram showing an example of the construction of a search assistant interface in a client.

5 Description of the Preferred Embodiments

Fig. 1 shows a typical general arrangement wherein a client 11 specifies an arbitrary set of key documents in a document database 131 of a server 13, and obtains a set of documents having a high relevance (similarity) to the specified set of key documents from a document database 141 of another server 14. Herein, the source and target document databases 131, 141 are located on servers in different places which can be respectively accessed via a network 12.

10 First, the client 11 specifies a set of key documents in the source document database 131 according to user's specification, and sends this information to the server 13 as a set comprising a document identifier, for example an ID attached to each document which the server 13 can understand, via the network 12. The set of documents is specified in a window for displaying search results P1 described later.

15 The server 13 identifies a set of documents which were sent from the client. A summary of the set of documents is then made for the searched set of documents by a summary making module 132, and this is sent back to the client 11 via the network 12. Herein, the term "summary" means a set of topic words relating

to a set of documents. The summary making module can be constructed by any of the known methods such as that disclosed in Japanese Patent Laid-Open No. Hei 9-62693, "Method of Document Classification by Probability Model".

As an example, word frequencies are first totaled by splitting up all the documents in the set of documents for which it is desired to make a summary, into words. In general, as the degree to which a given set of documents is represented by particular words is higher for words which appear more frequently within it, words will tend to be included in the summary more frequently the higher their occurrence frequency is in the set of documents. However, general words which often appear in all documents such as "do", etc., are not suitable for the summary. Therefore, words are usually selected for inclusion in the summary by considering also their appearance frequency in the document database to which the set of documents belongs. Specifically, desirable words are topic words having a high occurrence frequency in a specified set of documents but a low overall frequency in the document database overall, i.e., they are suitable for a summary characterizing the set of documents. Hence, words are selected for the summary by calculating their weighting from suitable parameters using the occurrence frequency in the set of documents and the occurrence frequency in the document database as input, and adopting words

having a weighting equal to or greater than a certain threshold.

The higher the weighting, the higher the relevance of the word to a given document, and the lower the weighting, the lower the relevance of the word to the document.

The server 13 returns a set of words having a weighting calculated by the above-mentioned method to the client via the network 12. These words are displayed as "topic words" in Fig. 2.

Next, at the client 11, users evaluate or process the summary (summary of the set of key documents) from the server 13, and the client 11 transmits it to the target server 14 via the network 12.

In the evaluation or the processing performed by users at the client, users for example remove words which are not deemed to be relevant from the summary, or users for example replace words in the summary.

Using the search module 143, the server 14 calculates the relevance of the summary of the set of key documents sent from the client to the target document database 141, and returns document identifiers of high relevance to the client 11 with a relevance weighting. The search module here can be implemented by a keyword search known in the art. Specifically, as the summary of the set of documents which is input is a set of words with weightings, these words may be considered as weighted input keywords and an OR keyword

search performed. In this case, the weighting (relevance) of the document which is the search result can be calculated. This is done by taking words which appear both in the summary and the document to be searched, calculating an overall weighting from their weighting in the summary and their weighting in the document to be searched (e.g., product of the two weightings), and then adding up the weightings of all the words (e.g., calculating a sum total) to obtain the relevance.

Using the above method, the client 11 can obtain a set of documents in the document database 141 which relates to an arbitrary set of key documents in the document database 131. The characteristic feature of this method is that network traffic is reduced to a small amount by leaving the processing (summary making) of the original set of documents searched to the server side. It will be appreciated that the amount of traffic is much less than in the case where the client has to receive and process all of the document information to be searched. The search assistant module 112 of the client then basically has only to send the summary of the set of documents from the source server to the target server, and almost all of the processing involved in the search can be left to both servers. Moreover, the server side merely has to have a summary making module and a search module for the document database in question, and it is therefore

completely unnecessary to consider information in other document databases.

In the aforesaid description, a method was described wherein the document database 131 was the source database and the document database 141 was the target database, but the same method can be adopted wherein the document database 141 is the source database and the document database 131 is the target database. In this case, the client obtains a summary of the set of key documents from a summary making module 142 of the server 14, transmits it to the server 13 which is to be searched, and obtains relevant documents in the document database 131 from the search module 133 of the server 13. If the above is generalized, and a server with a summary making module and a search module is provided for a new document database, this document database can be made to function as the source database or target database for all document databases connected to the network simply by connecting the server to the network.

In Fig. 1, the summary making module and search module (132, 133, and 142, 143) were respectively located in different servers (13,14), but this embodiment of the invention is not limited to this particular arrangement. For example, the summary making module and search module may be installed in a different server from the document database, and a summary making module and search module provided by

this server for plural databases.

Finally, Fig. 2 shows an embodiment concerning the client. 111 is an example of a search assistant interface installed in the client. This is basically the same as the interface proposed by the inventor of the present application in Japanese Patent Laid-Open No. Hei 11-85786, "Document search support method and document search support service" (corresponding to U.S. Patent Application S.N.09/145,155, filed 09/01/98 by Nishioka et al), or Japanese Patent Laid-Open Hei 10-74210, "Document search support method and device, and document search service using same", (corresponding to U.S. Patent Application S.N.08-888,017, filed 07/03/97 by Niwa et al). E1 is a window for inputting a search query, wherein the user can input a search query by a string of keywords or in the form of a sentence. M1 is a window for selecting a document database wherein the user can pull down a specific part on the right edge with a mouse to show a list of document databases, and select a desired document database. B1 is a search button which initiates a search. Therefore, the user inputs an arbitrary search query in the window E1, selects a document database to be searched in the window M1, and performs an ordinary search by keywords input to the window E1 concerning the document database selected in the window M1, by pressing the button B1. This search is performed with the support of the search assistant module 112 shown in Fig. 1, but as the

details of the search method were given in the previous application, they will not be repeated here.

P1 is a window for displaying a search result. In the upper part, a panel showing the total number of documents retrieved as a result of the search process and a number of documents selected by the user as described hereafter, is displayed. Underneath this, a panel is provided for the user to specify selected/not selected(P13), and a document title part showing the relevance(P12) to the search query and titles(P11) of documents displayed in the form of a list. This display window has a scroll function so that, by scrolling, the user can see a part which cannot be displayed in the display at one time. In the selected/not selected panel, documents are either selected or deselected each time there is a mouse click. When documents are selected by clicking, a summary of the corresponding documents is displayed as a graphical representation of a set of words with weightings in a summary display window P2. The summary display window P2 also has a panel in its upper part where the total number of topic words and the number of topic words selected by the user are displayed. Document titles are usually sorted in order of relevance.

The window P1 for displaying the search result in the diagram shows that a total of 22 documents were retrieved as a result of the search, and that three documents were selected by the user as interesting

documents judging from their titles. The selected documents are given a check mark by clicking. In the summary display window P2, five topic words are accordingly displayed corresponding to the selected documents.

Although omitted from this embodiment, conversely, documents for which the topic words selected in the summary display window P2 are representative, can be displayed in the window P1.

Therefore, the user can perform a more advanced search by making a summary customized according to his preference. This is explained in detail in the aforesaid reference Japanese patent laid open Hei 11-085786.

Hence, the user can select/deselect documents while referring to the titles and the topic words of selected documents, and can select plural documents in which he is interested.

Subsequently, if the user is interested in handling other document data for the set of documents corresponding to this search result, he may change the document database in the window M1, and press the button B1 to begin a new search.

Hence, the client sends an identifier of the plural documents selected to the server where the source document database is stored (for example, the server 13), obtains a summary of these plural documents, sends this summary to the server where the target

document database is stored (for example, the server 14), and obtains a search result from the target server (for example, the server 14). The new search result is displayed in the window P1. In other words, in this case, P1 is updated by the set of documents which was newly searched.

To compare a new search result with a previous search result, the user may press a back button B2 to re-display the previous search result in the window P1, and return the display of the window P1 to its state before search was performed. Likewise, the window P1 can be advanced to the new search result by pushing a forward button B3.

As the user can search other document databases corresponding to such a search result at any stage of the search, the user can freely proceed from one database to another database by repeating the search cycle. Naturally, it is also possible to repeat this cycle within the same document database, i.e., without changing the document database.

According to this invention, the user can freely specify a document database to be searched and freely enhance the search without concern for the location or composition of each document database. Further, as a server in which a document database is located can be modularized, the server can be made to function as an source database or a target database with respect to all other databases connected to a

network simply by connecting a server comprising a summary making module and a search module to the network when it is desired to search new document databases.

While the present invention has been described above in conjunction with the preferred embodiments, any person skilled in the art would be enabled by this disclosure to make various modifications to this embodiment and still be within the scope and spirit of the invention as defined in the appended claims.